Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 29 - Hypotheses testing. One-sample t-test and application to linear regression

Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
salvatore.ruggieri@unipi.it

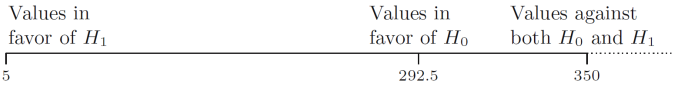
Hypothesis testing

- We tested how likely is *Exp*() as data generation model for the *software* dataset
- Hypotheses testing consists of contrasting two conflicting hypotheses based on observed data
- Consider the German tank problem:
  - Military intelligence states that $N = 350$ tanks were produced *[H0 or null hypothesis]*
  - Alternative hypothesis: *[H1 or alternative hypothesis]*
    $N < 350$ (*one-tailed or one-sided test*), or $N \neq 350$ (*two-tailed or two-sided test*)
  - Observed serial tank id's: 61 19 56 24 16
- Statistical test: How likely is the observed data under the null hypothesis?
  - If it is NOT (sufficiently) likely, we reject the null hypothesis in favor of H1
  - If it is (sufficiently) likely, we cannot reject the null hypothesis
- Why '*we cannot reject the null hypothesis*' and not instead '*we accept the null hypothesis*'?
  - Other hypotheses, e.g., $N = 349$ or $N = 351$, could also be not rejected
    and then, we cannot say which of $N = 349$ or $N = 350$ or $N = 351$ is actually true

Test statistic

TEST STATISTIC. Suppose the dataset is modeled as the realization of random variables $X_1, X_2, \ldots, X_n$. A *test statistic* is any sample statistic $T = h(X_1, X_2, \ldots, X_n)$, whose numerical value is used to decide whether we reject H_0.

- In the German tank example: *[See Lesson 19]*
  - $H_0 : N = 350$
  - $H_1 : N < 350$
  - Observed serial tank id's: 61 19 56 24 16
- We use $T = \max\{X_1, X_2, X_3, X_4, X_5\}$
- If H_0 is true, i.e., $N = 350$, then $E[T] = \frac{5}{6}(N+1) = \frac{5}{6}351 = 292.5$



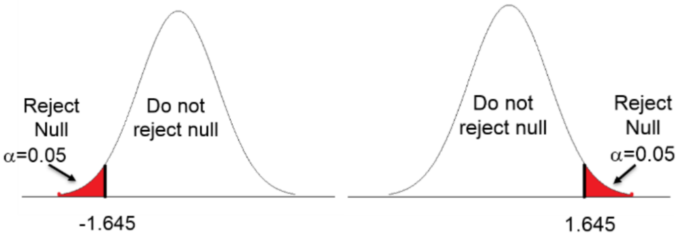
- If H_0 is true, we have:

$$P(T \leq 61) = P(\max\{X_1, X_2, X_3, X_4, X_5\} \leq 61) = \frac{61}{350} \cdot \frac{60}{349} \cdots \frac{57}{346} = 0.00014$$

very unlikely: either we are unfortunate, or H_0 can be rejected

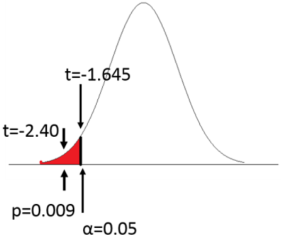
Statistical test of hypothesis: one-tailed – critical region

- H_0: $\theta = v$ *[Null hypothesis]*
- H_1: $\theta < v$ (resp. H_1: $\theta > v$) *[Left-tailed/Right-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T = h(X_1, \ldots, X_n)$ test statistics when H_0 is true
- $x_1, \ldots, x_n$: observed dataset, and $t = h(x_1, \ldots, x_n)$ *[t-value]*
- c_l s.t. $P(T \leq c_l) = \alpha$ (resp. c_u s.t. $P(T \geq c_u) = \alpha$) *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $t \leq c_l$ (resp. $t \geq c_u$): H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected



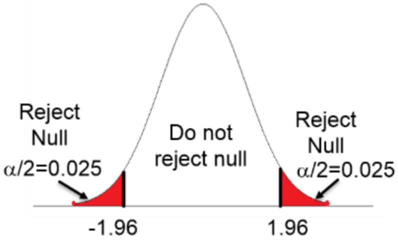
Statistical test of hypothesis: one-tailed – p-value

- H_0: $\theta = v$ *[Null hypothesis]*
- H_1: $\theta < v$ (resp. H_1: $\theta > v$) *[Left-tailed/Right-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T = h(X_1, \ldots, X_n)$ test statistics when H_0 is true
- $x_1, \ldots, x_n$: observed dataset, and $t = h(x_1, \ldots, x_n)$ *[t-value]*
- $p = P(T \leq t)$ (resp. $p = P(T \geq t)$) *[p-value]*
  - evidence against H_0 - the smaller the stronger evidence
- Output of the test at confidence level $100(1-\alpha)\%$ using p-values
  - $p \leq \alpha$: H_0 is rejected
  - otherwise: H_0 cannot be rejected



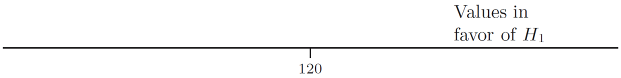
Statistical test of hypothesis: two-tailed

- H_0: $\theta = v$ *[Null hypothesis]*
- H_1: $\theta \neq v$ *[Two-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T = h(X_1, \ldots, X_n)$ test statistics when H_0 is true
- $x_1, \ldots, x_n$: observed dataset, and $t = h(x_1, \ldots, x_n)$ *[t-value]*
- c_l s.t. $P(T \leq c_l) = \alpha/2$ and c_u s.t. $P(T \geq c_u) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $t \leq c_l$ or $t \geq c_u$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected



Example: speed limit

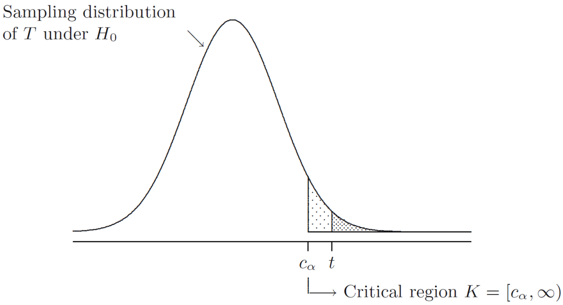
- Speed limit: 120 Km/h
- A device conduts 3 measurements: $X_1, X_2, X_3 \sim \mathcal{N}(\mu, 4)$ (true speed + measur. error)
- Based on $T = \bar{X}_3 = (X_1 + X_2 + X_3)/3 \sim \mathcal{N}(\mu, 4/3)$:
  - if $T > c_u$ the driver is fined
  - otherwise it is not
- What should c_u be to unjustly fine only 5% of drivers? *[Type I error]*
- One-tailed statistical test
  - H_0: $\mu = 120$ (null hypothesis)
  - H_1: $\mu > 120$ (alternative hypothesis)
  - $\alpha = 0.05$ (significance level), or $100(1 - \alpha)\% = 95\%$ (confidence level)
  - $T = \bar{X}_3$ (test statistics)
- Assuming H_0 is true, find t such that $P(T \geq c_u) = 0.05$



Example: speed limit

- $X_1, X_2, X_3 \sim \mathcal{N}(\mu, 4)$ and then $T = \bar{X}_3 \sim \mathcal{N}(\mu, 4/3)$
- $Z = \frac{T-120}{2/\sqrt{3}} \sim \mathcal{N}(0,1)$
- $P(T \geq c_u) = P(\frac{T_3-120}{2/\sqrt{3}} \geq \frac{c_u-120}{2/\sqrt{3}}) = P(Z \geq \frac{c_u-120}{2/\sqrt{3}})$
- Right critical value: $P(Z \geq z_\alpha) = \alpha$
- Hence $\frac{c_u-120}{2/\sqrt{3}} = z_{0.05}$, i.e., $c_u = 120 + z_{0.05}\frac{2}{\sqrt{3}} = 121.9$
- In summary, for $\alpha = 0.05$ we should reject $H_0 : \mu = 120$ in favor of $H_1 : \mu > 120$ if the observed (average) speed t is $t \geq 121.9$

Critical values and p-values



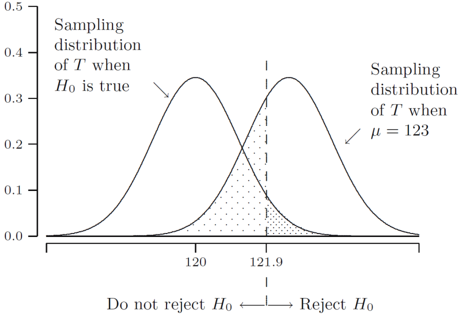
- *Critical region* K: the set of values that reject H_0 in favor of H_1 at significance level α
- *Critical values*: values on the boundary of the critical region
- *p-value*: the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that H_0 is true
- $t \in K$ iff p-value $\leq \alpha$

Type I and Type II errors

		True state of nature	
		H_0 is true	H_1 is true
Our decision on the basis of the data	Reject H_0	*Type I error*	Correct decision
	Not reject H_0	Correct decision	*Type II error*

- Type I error is we falsely reject H_0: $P(\text{Reject } H_0 | H_0 \text{ is true})$ *[α-risk, false positive rate]*
  - E.g., unjust speed-limit fine
  - we reject H_0 when $p < \alpha$, so this error occur with probability $100\alpha\%$
  - this error can be controlled by setting the significance level α to the largest acceptable value
    - how much is an *acceptable value*?
  - A possible solution is to solely report the p-value, which conveys the maximum amount of information and permits decision makers to choose their own level
- Type II error is we falsely do not reject H_0: $P(\text{Not Reject } H_0 | H_1 \text{ is true})$ *[β-risk, false negative rate]*
  - E.g.,lack of a true speed-limit sanction
  - $1 - \beta = P(\text{Reject } H_0 | H_1 \text{ is true})$ is called the *power* of the test

Type II error: how large can it be?



- Type II error: probability of not being fined when $\mu > 120$ but $t < 121.9$
- Assume $\mu = 125$, hence $T = \bar{X}_3 \sim \mathcal{N}(125, 4/3)$
  - Type II error is $P(T < 121.9|\mu = 125) = P(\frac{T-125}{2/\sqrt{3}} < \frac{121.9-125}{2/\sqrt{3}}) = \Phi(-2.68) = 0.0036$
- Assume $\mu = 123$, hence $T = \bar{X}_3 \sim \mathcal{N}(123, 4/3)$
  - Type II error is $P(T < 121.9|\mu = 123) = P(\frac{T-123}{2/\sqrt{3}} < \frac{121.9-123}{2/\sqrt{3}}) = \Phi(-0.95) = 0.1711$
- Type II error can be arbitrarily close to $1 - \alpha$

Relation with confidence intervals

- H_0: $\mu = 120$ (null hypothesis)
- H_1: $\mu > 120$ (alternative hypothesis)
- $\alpha = 0.05$ (significance level)
- $c_u = 120 + z_{0.05}\frac{2}{\sqrt{3}} = 121.9$
- H_0 rejected when:

$$
\begin{aligned}
& t = \bar{x}_3 \geq c_u \\
\Leftrightarrow \quad & \bar{x}_3 \geq 120 + z_{0.05}\frac{2}{\sqrt{3}} \\
\Leftrightarrow \quad & 120 \leq \bar{x}_3 - z_{0.05}\frac{2}{\sqrt{3}} \\
\Leftrightarrow \quad & \text{120 is not in the 95\% one-tailed c.i. for } \mu
\end{aligned}
$$

because $(\bar{x}_3 - z_{0.05}\frac{2}{\sqrt{3}}, \infty)$ is a one-tailed c.i. for μ

One sample tests for the mean: summary

- $x_1, \ldots, x_n$ realizations of $X_1, \ldots, X_n \sim F$ with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$

  Question: how consistent is the dataset with the null hypothesis that $\mu = \mu_0$

  - expected level over the population given blood measurement levels over n persons
  - expected accuracy over the distribution given results on n test instances for a classifier
- $H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$ (or $H_1 : \mu > \mu_0$, or $H_1 : \mu < \mu_0$)
- We distinguish a few cases:
  - Normal data $F = \mathcal{N}(\mu, \sigma^2)$
    - ☐ with known variance: $Z = \frac{\bar{X}_n - \mu_0}{\sigma / \sqrt{n}}$ *[z-test]*
    - ☐ with unknown variance: $T = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}}$ *[t-test]*
  - General data (with unknown variance)
    - ☐ large sample, i.e., large n, $T = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}}$ *[t-test]*
    - ☐ symmetric distribution *[Wilcoxon test]*
    - ☐ bootstrap t-test
  - Bernoulli data $F = Ber(\mu)$
    - ☐ Test of proportions : $B^\star = \frac{\bar{X}_n - \mu_0}{\sqrt{\mu_0(1-\mu_0)}/\sqrt{n}}$ *[Binomial test]*

Normal data with known σ^2: z-test

- $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$
- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$ *[Two-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ test statistics when H_0 is true
- $x_1, \ldots, x_n$: observed dataset, and z value is $\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$
- $P(Z \leq -z_{\alpha/2}) = \alpha/2$ and $P(Z \geq z_{\alpha/2}) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $|z| \geq z_{\alpha/2}$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected

See R script

Normal data with unknown σ^2: t-test

- $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$
- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$ *[Two-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t(n-1)$ test statistics when H_0 is true [recall $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$]
- $x_1, \dots, x_n$: observed dataset, and t value is $\frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$
- $P(T \leq -t_{\alpha/2,n-1}) = \alpha/2$ and $P(T \geq t_{\alpha/2,n-1}) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $|t| \geq t_{\alpha/2,n-1}$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected

See R script

General data, large sample: t-test

- $T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \to \mathcal{N}(0,1)$ for $n \to \infty$ *[Variant of CLT]*
- We can use z-test with $\sigma^2 = s_n^2$
- Or, since $t(n) \to \mathcal{N}(0,1)$ for $n \to \infty$, we can use t-test directly!

See R script

General data, symmetric distribution: Wilcoxon signed-rank test

- $X_1, \dots, X_n \sim F$ with $f(\mu - x) = f(\mu + x)$ (symmetric distribution)
- $H_0 : \mu = 67$
- $H_1 : \mu \neq 67$
- $W = \min \{\sum rank^+, \sum rank^-\}$, with ranking w.r.t. $|x_i - \mu_0|$

x	71	79	40	70	82	72	60	76	69	75
$x - \mu_0$	4	12	-27	3	15	5	-7	9	2	8
$rank$	3	8	10	2	9	4	5	7	1	6
$rank^+$	3	8		2	9	4		7	1	6
$rank^-$			10				5			

- $w = \min \{40, 15\} = 15$
- Ignore cases where $|x_i - \mu_0| = 0$. If the values have ties, then consider the mean value
- Normal approximation for $n > 50$
- Exact test for $n \leq 50$ *[see the* **null distribution***]*
- Also, a statistical test of the median (for symmetric distributions)!

General data: bootstrap test (see Lesson 27)

`boot.ci` method in R confidence intervals:

- `type='stud'`: $\left(\bar{x}_n - q_{1-\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n - q_{\alpha/2}\frac{s_n}{\sqrt{n}}\right)$ with quantiles over the distribution of t^*

EMPIRICAL BOOTSTRAP SIMULATION FOR THE STUDENTIZED MEAN. Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F. The expectation corresponding to F_n is $\mu^* = \bar{x}_n$.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.
2. Compute the studentized mean for the bootstrap dataset:

$$t^* = \frac{\bar{x}_n^* - \bar{x}_n}{s_n^*/\sqrt{n}},$$

where $\bar{x}_n^*$ and s_n^* are the sample mean and sample standard deviation of $x_1^*, x_2^*, \ldots, x_n^*$.

Repeat steps 1 and 2 many times.

- $t_0 = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$ $\quad r$ number of repetitions
- one-sided p-value, i.e., $P(T \geq t_0)$, estimated as $|\{i = 1, \ldots, r \mid t_i^* \geq t_0\}|/r$
- two-sided p-value, i.e., $P(|T| \geq |t_0|)$, estimated as $|\{i = 1, \ldots, r \mid |t_i^*| \geq |t_0|\}|/r$

See R script

Hypothesis testing for a proportion: the binomial test

- Dataset $x_1, \ldots, x_n$ realization of $X_1, \ldots, X_n \sim Ber(\theta)$
- $H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0$
- Test statistics: $B = \sum_{i=1}^n X_i \sim Bin(n, \theta_0)$ *[Asymmetric distribution]*
- b-value is $\sum_{i=1}^n x_i$
- Critical values (exact test):

$$P(B \leq l) = \sum_{i=0}^{l} \binom{n}{i} \theta_0^i (1-\theta_0)^{n-1} = P(B \geq u) = \sum_{i=u}^{n} \binom{n}{i} \theta_0^i (1-\theta_0)^{n-i} = \alpha/2$$

- Normal approximation $Bin(n, \theta_0) \approx \mathcal{N}(n\theta_0, n\theta_0(1-\theta_0))$
  - scaled test statistics:

$$B^\star = \frac{B - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \sim \mathcal{N}(0,1)$$

  - use z-test with $\sigma^2 = \theta_0(1-\theta_0)$ because $B^\star = \frac{B/n - \theta_0}{\sqrt{\theta_0(1-\theta_0)}/\sqrt{n}} = \frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}}$
  - or even t-test for large samples

See R script

Hypothesis testing in linear regression

- Simple linear regression: $Y_i = \alpha + \beta x_i + U_i$ with $\underline{U_i \sim \mathcal{N}(0, \sigma^2)}$
- We have $\hat{\beta} \sim \mathcal{N}(\beta, Var(\hat{\beta}))$ where $Var(\hat{\beta}) = \sigma^2 / SXX$ is unknown *[see Lesson 20]*
- The studentized statistics is $t(n-2)$-distributed: *[proof omitted]*

$$T = \frac{\hat{\beta} - \beta}{\sqrt{Var(\hat{\beta})}} \sim t(n-2)$$

- $H_0 : \beta = 0 \qquad H_1 : \beta \neq 0$
- p-value is $p = P(|T| > |t|) = 2 \cdot P(T > \left|\frac{\hat{\beta}-0}{se(\hat{\beta})}\right|)$
- H_0 can be rejected in favor of H_1 at $\alpha = 0.05$, if $p < 0.05$, or, equivalently, if $|t| > t_{n-2,0.025}$.
- A similar approach applies to the intercept.

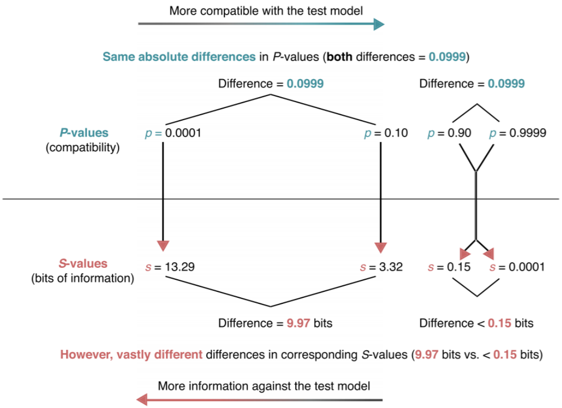
See R script

Misues of p-values

Misinterpretations of p-values, [Greenland et al, 2016]

- ~~The p-value is the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false.~~ A p-value indicates the degree of compatibility between a dataset and a particular hypothetical explanation
- ~~The 0.05 significance level is the one to be used:~~ No, it is merely a convention. There is no reason to consider results on opposite sides of any threshold as qualitatively different.
- ~~A large p-value is evidence in favor of the test hypothesis:~~ A p-value cannot be said to favor the test hypothesis except in relation to those hypotheses with smaller p-values
- ~~If you reject the test hypothesis because $p \leq 0.05$, the chance you are in error is 5%:~~ No, the chance is either 100% or 0%. The 5% refers only to how often you would reject it, and therefore be in error.

s-values



- Shannon information value or surprisal value (*s*-**value**) is $-\log_2 p$ (unit measure: bit)
  - $p = 0.5 \Rightarrow s = 1$ surprising as getting one heads on 1 fair coin toss
  - 9.97 bits difference surprising as getting all heads on 10 fair coin tosses

Optional references

- On confidence intervals and statistical tests (with R code)

Myles Hollander, Douglas A. Wolfe, and Eric Chicken (2014)
Nonparametric Statistical Methods.
3rd edition, *John Wiley & Sons, Inc.*

- On p-values

Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman (2016)
Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.
European Journal of Epidemiology 31, pages 337–350